Random generation of optimal saturated designs

An approach based on discovery probability

Roberto Fontana

Received: date / Accepted: date

Abstract Efficient algorithms for searching for optimal saturated designs are widely available. They maximize a given efficiency measure (such as D-optimality) and provide an optimum design. Nevertheless, they do not guarantee a *global* optimal design. Indeed, they start from an initial random design and find a local optimal design. If the initial design is changed the optimum found will, in general, be different. A natural question arises. Should we stop at the design found or should we run the algorithm again in search of a better design? This paper uses very recent methods and software for discovery probability to support the decision to continue or stop the sampling. A software tool written in SAS has been developed.

Keywords Design of experiments · Optimal designs · Unobserved species · Discovery probability

1 Introduction

In the design of experiments, optimal designs, or optimum designs, are a class of experimental designs that are optimal with respect to a given statistical criterion.

In this paper we focus on saturated optimum designs (SOD). Saturated designs contain a number of points that is equal to the number of parameters of the model. It follows that SODs are often used in place of standard designs, such as orthogonal fractional factorial designs, when the cost of each experimental run is high. Main references to this topic include Atkinson et al (2007), Pukelsheim (2006), Shah and Sinha (1989) and Wynn (1970).

The optimality of a design depends on the statistical model that is assumed and is assessed with respect to a statistical criterion, which, for information-based criteria, is related to the variance-matrix of the model parameter estimators. Well-known and commonly used criteria are A-optimality and D-optimality.

Widely used statistical systems like SAS and R have procedures for finding an optimal design according to the user's specifications. In this paper we will refer to Proc Optex of SAS/QC (sas (2010)), but the approach can be adopted for other software.

The Optex procedure searches for optimal experimental designs. The user specifies an efficiency criterion, a set of candidate design points, a linear model and the size of the design to be found and the procedure generates a subset of the candidate set so that the terms in the model can be estimated as efficiently as possible. By default, the standard output of the procedure is a list of 10 designs that are found as the result of 10 runs of the exchange search algorithm (Mitchell and Miller Jr (1970)) starting each time from an initial completely randomly chosen design.

The number of times that we decide to run the search algorithm is crucial. Obviously, if we increase it, in general we will explore different local optima with the possility to find better designs. On the other hand, sometimes, the extra time that we use to explore other possibilities is wasted because new optima do not exist. This work aims at developing a methodology that could support the user in making the decision whether to stop or continue the search.

The paper is organized as follows. In Sect. 2 we state the problem of finding new optimal designs as the problem of finding new species in a population. Then, in Sect. 3, using some examples, we describe how our methodology, which is based on the estimator of the

R. Fontana
Department of Mathematical Sciences - Politecnico di Torino
Tel.: +39-011-0907504 E-mail: roberto.fontana@polito.it

discovery probability, could be used for optimal design generation. In Sect. 4 we describe the algorithm in more detail. The software code that has been developed is written in SAS, is available on request and can be used for any choice of factors, levels and model. Concluding remarks are in Sect. 5.

2 Optimal designs vs richness of species

We consider the following setting that is quite common in optimal design problems.

We have d factors, $A_1, \ldots, A_d$. The factor A_i has s_i levels coded with the integer $0, \ldots, s_i - 1$, $i = 1, \ldots, d$. The full factorial design is $\mathcal{D} = \{0, \ldots, s_1 - 1\} \times \ldots \times \{0, \ldots, s_m - 1\}$. For each point $\zeta = (\zeta_1, \ldots, \zeta_d)$ of $\mathcal{D}$ we consider a real-valued random variable $Y_{\zeta_1,\ldots,\zeta_d}$. We make the hypothesis that the means of the responses, $E\left[Y\right]$ where Y is the column vector $[Y_\zeta; \zeta \in \mathcal{D}]$ can be modeled as

$$E\left[Y\right] = X_{\mathcal{D}}\beta\,, \tag{1}$$

where $X_{\mathcal{D}}$ is the non-overparametrized design matrix, as it will be defined in Sect. 2.1, and β is the subset of all the effects (constant effect, main effects and interactions) that are supposed to affect the response Y.

Given an efficiency criterion ϕ, a saturated optimal design (ϕ-SOD) is a subset of the full factorial design $\mathcal{D} = \{0, \ldots, s_1 - 1\} \times \ldots \times \{0, \ldots, s_m - 1\}$, whose size is equal to the number of degrees of freedom of the model (1) and that maximizes this criterion ϕ. In this paper we focus on information-based criteria and, in particular, on D-optimality but other criteria can be chosen (like A-optimality and G-optimality). We denote this type of problem with the triple $(\mathcal{D}, \mathcal{M}, \phi)$ where $\mathcal{D}$ is the full design, $\mathcal{M}$ is the hypothesized model (see Eq. 1) and ϕ is the optimality criterion.

Given a subset $\mathcal{F}$ of $\mathcal{D}$, the information matrix is defined as $X'_{\mathcal{F}}X_{\mathcal{F}}$ where $X_{\mathcal{F}}$ is the design matrix corresponding to $\mathcal{F}$ and X' is the transpose of X. D-optimality aims at maximizing $D_{\mathcal{F}}$, the determinant of the information matrix

$$D_{\mathcal{F}} = \det(X'_{\mathcal{F}}X_{\mathcal{F}})\,. \tag{2}$$

There are several algorithms for searching for D-optimal designs. They have a common structure. They start from an initial design, randomly generated or user specified, and move, in a finite number of steps, to a better design. In general, if a different initial design is chosen, a different optimal design is found.

It follows that, given an algorithm α, a population $\mathcal{A}^D_\alpha$ of D-optimal designs can be defined. This population is made up of all the saturated designs that are the result of the execution of the algorithm α and is a subset of all the subsets of $\mathcal{D}$ of size equal to the number of degrees of freedom of the model.

The elements of $\mathcal{A}^D_\alpha$ can be classified into species, according to the criterion for which $\mathcal{F}_1 \in \mathcal{A}^D_\alpha$ and $\mathcal{F}_2 \in \mathcal{A}^D_\alpha$ are of the same species if and only if they have the same value in terms of the D criterion, $D_{\mathcal{F}_1} = D_{\mathcal{F}_2}$.

We observe that, as proved in Proposition 1, isomorphic designs belong to the same species, while, in general, the opposite is not true because there are designs with the same value of the D criterion but that are not isomorphic. As is known two designs are isomorphic if one can be obtained from the other by relabeling the factors, reordering the runs, and switching the levels of factors, e.g. Clark and Dean (2001).

Proposition 1 *Let us consider $\mathcal{F}_1 \subseteq \mathcal{D}$ and $\mathcal{F}_2 \subseteq \mathcal{D}$. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are isomorphic then $D_{\mathcal{F}_1} = D_{\mathcal{F}_2}$.*

Proof We separately analyse row/column permutations and the switching of the levels of some factors. If $\mathcal{F}_2$ is obtained permuting the rows and/or the columns of $\mathcal{F}_1$ it follows that

$$X_{\mathcal{F}_2} = RX_{\mathcal{F}_1}C$$

where R and C are permutation matrices. Then

$$D_{\mathcal{F}_2} = \\ = \det((X'_{\mathcal{F}_2}X_{\mathcal{F}_2})) = (\det(R))^2 \det((X'_{\mathcal{F}_1}X_{\mathcal{F}_1}))(\det(C))^2 = \\ = D_{\mathcal{F}_1}$$

being $\det(R) = \det(C) = 1$. A similar argument holds for switching the levels of some factors. □

Studying the species of $\mathcal{A}^D_\alpha$ or, in general, of $\mathcal{A}^\phi_\alpha$ where ϕ is an optimal criterion, is interesting for optimal design generation. Let us consider the problem $(\mathcal{D}, \mathcal{M}, \phi)$ and let us choose an algorith α to search for ϕ-SODs. If we run this algorithm n times, each time starting from a completely random initial design, we will get a sample of n elements of $\mathcal{A}^\phi_\alpha$. Such elements can be classified in $k_n \leq n$ different species according to the value of the criterion ϕ. Recent methods for discovery probability estimation, Favaro et al (2012), can be applied to the vector $(\ell_1, l_2, \ldots, \ell_n)$ where ℓ_r is the number of species in the sample with frequency r, $r = 1, \ldots, n$. In particular, based on a sample of size n, for any additional unobserved sample size $m \geq 0$ and for any frequency $k = 0, \ldots, n+m$, these methods provide, an explicit estimator for the probability $U_{n+m}(k)$ that the $(n + m + 1)$-th observation coincides with a species whose frequency, within the sample size $n + m$, is exactly k. The case $m = k = 0$ corresponds to assessing the probability of finding a new species in the subsequent observation, that in the context of optimal designs, is the probability of finding a saturated design with a different value of the criterion ϕ in the subsequent run of the algorithm. If this probability $U_{n+0}(0)$ is sufficiently high (let us say greater than 0.1 or even 0.05) it would be convenient to run the algorithm again because it is likely that we could find a new optimal design. If we found a new design, it could have a greater value of ϕ and this obviously represents an improvement to our optimization process. Even if this new design did not have an higher value of ϕ than the existing ones, this would give the possibility to increase the known part of $\mathcal{A}_\alpha^\phi$. In particular, for D-optimal designs, from Proposition 1, we know that designs with different values of $D_{\mathcal{F}}$ are non-isomorphic designs. It is quite common, in practical applications, to choose a design where the optimal criterion has a slightly smaller value than the maximum obtained but which has other better characteristics, such as space filling properties. The knowledge of a set of non-isomorphic designs can also be used for non parametric testing procedures, Giancristofaro et al (2012) and Basso et al (2004).

2.1 The design matrix

The design matrix $X_{\mathcal{D}}$ in Eq. 1 is built as follows.

- The first column is equal to 1 and corresponds to the constant effect, denoted by μ. The constant effect is always considered as a term of the model.
- If the main effect of the factor A_i is to be considered in the model, the corresponding $s_i - 1$ columns are computed as follows. For a design point with A_i at its k-th level
  - if $1 \leq k \leq s_i - 1$ the columns are all 0 except for the k-th column that is 1;
  - if $k = s_i$ the columns are all -1
- If an interaction $A_{i_1} \star \ldots \star A_{i_k}$ is to be considered in the model, the corresponding $(s_{i_1}-1)\cdot \ldots \cdot (s_{i_k}-1)$ columns are computed by taking the horizontal direct product of the colummns corresponding to the main effects of $A_{i_1}, \ldots, A_{i_k}$.

This coding corresponds to modeling without over parametrization and $X_{\mathcal{D}}$ is full rank.

For a subset $\mathcal{F}$ of $\mathcal{D}$, the design matrix $X_{\mathcal{F}}$ is simply built deleting from $X_{\mathcal{D}}$ the rows that correspond to the points of $\mathcal{D}$ that are not in $\mathcal{F}$.

2.2 Discovery probability

We briefly summarize the main results that are used in this work, as in Favaro et al (2012). The interested reader should refer to the original paper for a detailed description of the methodology.

Given a sample of size n, $(\ell_1, \ldots, \ell_n)$, where ℓ_r is the frequency of species that have been observed r-times in the sample, $r = 1, \ldots, n$. We have $\sum_{i=1}^n i\ell_i = n$. We denote the number of different species that have been observed in the sample by j. We get $\sum_{i=1}^n \ell_i = j$.

Based on a sample of size n, for an additional unobserved sample size $m \geq 0$ and for any frequency $k = 0, \ldots, n+m$, using a non parametric Bayesian approach, Favaro et al provide an estimator for the probability U_{n+m}^k that the $(n+m+1)$-th observation coincides with a species whose frequency, within the sample of size $n+m$, is exactly k.

We are interested in discovering new species, that correspond to the case $k = 0$.

From Section 2 of on p.1190 we obtain

$$U_{n+0}(0) = \frac{V_{n+1,j+1}}{V_{n,j}}$$

where, for the two-parameter Poisson-Dirichlet process, we have $V_{n,j} = \prod_{i=1}^{j-1}(\theta + i\sigma)/(\theta+1)_{n-1}$, $\sigma \in (0,1)$, $\theta > -\sigma$. The symbol $(a)_n$ denotes the n-th ascending factorial of a, $(a)_n = a(a+1)\ldots(a+n-1)$, $(a)_0 \equiv 1$. It follows that

$$U_{n+0}(0) = \frac{\theta + j\sigma}{\theta + n}$$

and, for $m > 0$, we obtain

$$U_{n+m}(0) = \frac{\theta + j\sigma}{\theta + n} \frac{(\theta + n + \sigma)_m}{(\theta + n + 1)_m}.$$

The estimates $\hat{\sigma}, \hat{\theta}$ of σ, θ are obtained as

$$\arg\max_{(\sigma,\theta)} \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} n! \prod_{i=1}^{n} \{\frac{(1-\sigma)_{i-1}}{i!}\}^{\ell_i} \frac{1}{\ell_i!}. \tag{3}$$

Using $(\hat{\theta}, \hat{\sigma})$ we finally obtain the estimates of the discovery probability at the $(n+1)$-th observation

$$\hat{U}_{n+0}(0) = \frac{\hat{\theta} + j\hat{\sigma}}{\hat{\theta} + n} \tag{4}$$

and at the $(n+m+1)$-th observation, $m > 0$,

$$\hat{U}_{n+m}(0) = \frac{\hat{\theta} + j\hat{\sigma}}{\hat{\theta} + n} \frac{(\hat{\theta} + n + \hat{\sigma})_m}{(\hat{\theta} + n + 1)_m} \tag{5}$$

3 Methodology and Applications

We repeat the search for optimal designs to analyse the population $\mathcal{A}_{\alpha}^{D}$ of D-optimal designs that can be found for a given problem using a predefined algorithm. Each time the algorithm starts from a randomly chosen initial design. We set a maximum number of iterations equal to $M_{\star}$ and we continue the process until the estimate of the discovery probability at the subsequent observation goes under a given threshold $p_{\star}$ or the maximum number of iterations is reached.

The procedure can be described as follows. A problem $(\mathcal{D}, \mathcal{M}, \phi)$, with $\phi = D$ in our examples, is defined and an algorithm α for ϕ-optimal design generation is chosen. For each iteration s, $s = 1, \ldots, M_{\star}$,

1. using the algorithm α, a ϕ-optimal saturated design $\mathcal{F}_s$ is obtained;
2. the values of the ϕ-criterion of $\mathcal{F}_s$ is computed;
3. the vector $(\ell_1, \ldots, \ell_s)$ is built, where ℓ_r is the number of species with frequency r, $r = 1, \ldots, s$;
4. an estimate $(\hat{\sigma}_s, \hat{\theta}_s)$ is obtained, see Eq. 3;
5. an estimate of $\hat{U}_{s+0}(0)$ is computed using Eq. 4;
6. if $\hat{U}_{s+0}(0) < p_{\star}$ the algorithm stops, otherwise the next iteration $s+1$ is performed (if $s+1 > M_{\star}$ the algorithm stops).

The main output of the algoritm is a set of designs, where each design belongs to a different species, i.e. has a different value of the ϕ-criterion.

We show how the methodology works using the following problem. Let us consider 7 factors, each with 2 levels and the model that contains the overall mean, the main effects and all the 2-factor interactions for a total of $1 + 7 + 21 = 29$ degrees of freedom. We search for *saturated* D-optimal designs that is D-optimal designs that contains 29 points.

We use Proc Optex sas (2010) with the exchange method, which is its default search method. With the default setting, the algorithm starts from 10 initial randomly chosen designs providing 10 D-optimal designs. We consider the design with the highest value of the D-efficiency of the 10 optimal designs as the optimal design found by the algorithm.

Setting the seed that is used for the random generation of the initial designs at 6789, the best of the 10 optimal designs, that we denote by $\mathcal{F}_1$, has $D_{\mathcal{F}_1} = 9.0911E39$ and $E_{\mathcal{F}_1}^{D} = 82.3162$, where $E_{\mathcal{F}}^{D}$, the D-efficiency of a $\mathcal{F}$, is defined as

$$E_{\mathcal{F}}^{D} = 100 \times \left(\frac{1}{\#\mathcal{F}} D_{\mathcal{F}}^{\frac{1}{\#\mathcal{F}}} \right)$$

where $\#\mathcal{F}$ is the number of runs of $\mathcal{F}$ that coincides with the degrees of freedom of the model for saturated designs.

Table 1 Number ℓ_r of D optimal designs that have found r times, $r = 1, \ldots, 493$; only $\ell_r \neq 0$ are shown.

r	ℓ_r
1	47
2	18
3	7
4	10
5	2
6	4
9	2
11	1
12	1
14	2
15	1
16	1
17	2
20	1
36	1
39	1
40	1
46	1
Total	103

Now we run the procedure above with $M_{\star} = 1,000$ and $p_{\star} = 0.10$.

After 493 runs, the estimate of the discovery probability at the next observation becomes lower than $p_{\star} = 0.10$ and the algorithm stops ($\tilde{U}_{493+0}(0) \approx 0.099$). We find 103 different local D-optimal designs. All these designs are not isomorphic (Proposition 1). The maximum (minimum) value of D-efficiency is 85.6265 (78.9605).

We decide to continue the search for new species choosing $p_{\star} = 0.05$ and $M_{\star} = 2,000$. The latter value is chosen taking into account that using Eq. 5 we get $\tilde{U}_{493+1000}(0) = 0.049$ and $\tilde{U}_{493+2000}(0) = 0.035$. We observe that these supplementary runs are added to the previous ones.

After 1,271 supplementary runs the estimate of the discovery probability at the next observation becomes lower than 0.05, $\tilde{U}_{1764+0}(0) \approx 0.0499$. After $1,271 + 493 = 1,764$ simulations we observe 191 different D-optimal designs. The maximum value of D-efficiency is still 85.6265, while the minimum is 78.1134.

We can now use the Fedorov algorithm, Fedorov (1972), that is considered more reliable, even if slower, than the exchange algorithm. We keep the standard setting for which, at each iteration, 10 optimal designs are generated and the one among them that has the highest D-efficiency value is taken as the optimal design.

We choose 3456 as the initial seed. The first iteration provides an optimal design $\mathcal{F}_1$ with $E_{\mathcal{F}_1}^{D} = 82.7079$. Now we repeat the procedure with $M_{\star} = 1,000$ and $p_{\star} = 0.10$. After only 18 iterations, as $\tilde{U}_{18+0}(0) \approx 0.087$, the algorithm stops, with 4 different designs. The maximum (minimum) value of D-efficiency is 83.9844

(82.4212). We have empirical evidence that the Fedorov algorithm is more stable than the exchange algorithm. We observe that the best design found with the exchange algorithm, that has D-efficiency equal to 85.6265, is not found in this first sample. We were able to find it running the algorithm again with $M_\star = 1,000$ and $p_\star = 0.01$.

4 The algorithm

In this section we provide a detailed description of the algorithm that has been developed to study the population $\mathcal{A}_\alpha^D$ that contains all the D-optimal designs that can be found by the algorithm α.

A problem $(\mathcal{D}, \mathcal{M}, \phi = D)$ is defined and an algorithm α for D-optimal design generation is chosen. The set of candidates that, in our setting, is the full factorial design is generated using an ad-hoc module written in SAS/IML. The algorithm α can be chosen from a list of methods that includes the exchange algorithm and the Fedorov algorithm.

For each iteration s, $s = 1, \ldots, M_\star$,

1. using the algorithm α, a D-optimal saturated design $\mathcal{F}_s$ is obtained;
2. the value of the D-efficiency, $E^D_{\mathcal{F}_s}$, of $\mathcal{F}_s$ is computed;
3. the vector $(\ell_1, \ldots, \ell_s)$ is built, where ℓ_r is the number of species with frequency r, $r = 1, \ldots, s$;
4. an estimate $(\hat{\sigma}_s, \hat{\theta}_s)$ is obtained, see Eq. 3;
5. an estimate of $\hat{U}_{s+0}(0)$ is computed using Eq. 4;
6. if $\hat{U}_{s+0}(0) < p_\star$ the algorithm stops, otherwise the next iteration $s+1$ is performed (if $s+1 > M_\star$ the algorithm stops).

The main output of the algoritm is a set of designs, where each design belongs to a different species, i.e. has a different value of the D-criterion.

4.1 Steps 1 and 2

At iteration s, with the chosen algorithm α, the Proc Optex procedure is used to generate a D-optimal design, $\mathcal{F}_s$. The species of $\mathcal{F}_s$ is the value of its D-efficiency, $E^D_{\mathcal{F}_s}$. The value of the efficiency is rounded to four decimal digits to avoid creating different species from numerical effects.

4.2 Step 3

Using all the designs $\mathcal{F}_1, \ldots, \mathcal{F}_s$ with their corresponding D-efficiencies, $E^D_{\mathcal{F}_1}, \ldots, E^D_{\mathcal{F}_s}$ the vector $(\ell_1, \ldots, \ell_s)$ is built, where ℓ_r is the number of species with frequency r, $r = 1, \ldots, s$.

4.3 Step 4

An estimate $(\hat{\sigma}_s, \hat{\theta}_s)$ must be obtained searching for (σ, θ), $\sigma \in (0,1)$, $\theta > -\sigma$ that maximizes $\mathrm{f}(\sigma, \theta)$, (see Eq. 3),

$$\mathrm{f}(\sigma,\theta) = \frac{\prod_{i=1}^{j-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} n! \prod_{i=1}^{n} \{\frac{(1-\sigma)_{i-1}}{i!}\}^{\ell_i} \frac{1}{\ell_i!}$$

The Genetic Algorithm module of SAS/IML has been used. In order to manage the constraints $\sigma \in (0,1)$, $\theta > -\sigma$ the search has been performed in the region $\mathcal{R} = [\delta, 1-\delta] \times [-(1-\delta), T_M]$ with $\delta = 0.01$ and $T_M = 1,000$. This region contains the non-feasible region made by the points inside the simplex $\mathcal{S} = \mathcal{R} \cap \{(\sigma, \theta) : \theta \le -\sigma\}$ whose vertices are $(\delta, -(1-\delta))$, $(\delta, -\delta)$ and $(1-\delta, -(1-\delta))$. We observe that the edges of $\mathcal{S}$ contain non-feasible points.

We decided to manage this constraint with the penalty method, because this method usually works well when most of the points in the solution space do not violate the constraints, as in our problem. The way in which the penalty in the objective function for unsatisfied constraints has been imposed is described here.

From the point of view of the search of the point $(\sigma_\star, \theta_\star)$ that maximizes $f(\sigma, \theta)$, it is equivalent to consider $\log \mathrm{f}(\sigma, \theta)$ instead of $\mathrm{f}(\sigma, \theta)$

$$\log \mathrm{f}(\sigma,\theta) = \log(\prod_{i=1}^{j-1}(\theta + i\sigma)) + \log(n!) +$$

$$-\log((\theta+1)_{n-1}) + \log(\prod_{i=1}^{n}\{\frac{(1-\sigma)_{i-1}}{i!}\}^{\ell_i}) - \log(\ell_i!) .$$

Omitting the terms that do not depend on σ and θ and as $(a)_n = \frac{\Gamma(a+n)}{\Gamma(a)}$ where Γ is the gamma function, the previous equation becomes the function $\mathrm{f}_\star(\sigma, \theta)$ here

$$\mathrm{f}_\star(\sigma,\theta) = \mathrm{f}_\star^{(1)}(\sigma,\theta) + \mathrm{f}_\star^{(2)}(\sigma,\theta) ,$$

where

$$\mathrm{f}_\star^{(1)}(\sigma,\theta) = \sum_{i=1}^{j-1} \mathrm{f}_\star^{(1,i)}(\sigma,\theta)$$

with $\mathrm{f}_\star^{(1,i)}(\sigma,\theta) = \log(\theta + i\sigma)$ and

$$\mathrm{f}_\star^{(2)}(\sigma,\theta) = -\log\Gamma(\theta+n) + \log\Gamma(\theta+1) +$$

$$+ \sum_{i=1}^{n} \ell_i \log\Gamma(i-\sigma) - j\log\Gamma(1-\sigma) .$$

We observe that, if the point $(\sigma, \theta) \in \mathcal{R}$ does not satisfy the constraint $\theta > -\sigma$ only $\mathrm{f}_\star^{(1)}(\sigma, \theta)$ becomes not defined. We apply a penalty value to $\mathrm{f}_\star^{(1)}(\sigma, \theta)$ and to $\mathrm{f}_\star^{(2)}(\sigma, \theta)$ as described below.

Given a point P_1 in the non-feasible region, $P_1 = (\sigma, \theta) \in \mathcal{S}$, $\tilde{P}_1$, the closest point to P_1 with respect to the euclidean distance that lies in the feasible region, is determined

$$\tilde{P}_1 = (\tilde{\sigma}, \tilde{\theta}) = (\frac{1}{2}(\sigma - \theta + \epsilon), \frac{1}{2}(\theta - \sigma + \epsilon))$$

where ϵ is a very small number to ensure that $\tilde{P}_1$ is feasible, i.e. $\tilde{P}_1 \in \mathcal{R} \cap \overline{\mathcal{S}}$. We used $\epsilon = 0.001$. The value of the function $\mathrm{f}_\star^{(1,1)}$ is computed in $\tilde{P}_1$ getting $\tilde{Y}_1 = \mathrm{f}_\star^{(1,1)}(\tilde{\sigma}, \tilde{\theta}) = \log \epsilon$. Then the value Y_1 of $\mathrm{f}_\star^{(1,1)}$ in P_1 is defined as $\mathrm{f}_\star^{(1,1)}(\sigma, \theta) = (1 + b_1)\tilde{Y}_1$ where b_1 is the euclidean distance between P_1 and $\tilde{P}_1$, $b_1 = \sqrt{\frac{1}{2}(\sigma + \theta - \epsilon)^2}$. In an analogous way, we apply this penalty method to all $P_i = (i\sigma, \theta)$ that eventually fall in the non-feasible region $\mathcal{S}$ getting $\mathrm{f}_{\star,P}^{(1)}(\sigma, \theta)$, the penalized version of $\mathrm{f}_\star^{(1)}(\sigma, \theta)$,

$$\mathrm{f}_{\star,P}^{(1)}(\sigma, \theta) = \sum_{i=1}^{j-1} \mathrm{f}_\star^{(1,i)}(\sigma, \theta)$$

where

$$\mathrm{f}_\star^{(1,i)} = \begin{cases} \log(\theta + i\sigma) & \text{if } \theta + i\sigma > 0 \\ (1 + b_i)\log(\epsilon) & \text{if } \theta + i\sigma \leq 0 \end{cases}, \; i = 1, \ldots, j{-}1 \,,$$

and b_i is the euclidean distance between $P_i = (i\sigma, \theta)$ and $\tilde{P}_i = (\frac{1}{2}(i\sigma - \theta + \epsilon, \frac{1}{2}(\theta - i\sigma + \epsilon)$ determined as described above. The penalized version $\mathrm{f}_{\star,P}^{(2)}(\sigma, \theta)$ of $\mathrm{f}_\star^{(2)}(\sigma, \theta)$ is simply defined as

$$\mathrm{f}_{\star,P}^{(2)}(\sigma, \theta) = \begin{cases} \mathrm{f}_\star^{(2)}(\sigma, \theta) & \text{if } \theta + \sigma > 0 \\ (1 + b_1)\,\mathrm{f}_\star^{(2)}(\sigma, \theta) & \text{if } \theta + i\sigma \leq 0 \\ & \text{ and } \mathrm{f}_\star^{(2)}(\sigma, \theta) \leq 0 \\ (1 - b_1)\,\mathrm{f}_\star^{(2)}(\sigma, \theta) & \text{if } \theta + i\sigma \leq 0 \\ & \text{ and } \mathrm{f}_\star^{(2)}(\sigma, \theta) > 0 \end{cases} .$$

We observe that

1. $p < q \Rightarrow b_p > b_q \; p, q = 1, \ldots, j - 1$;
2. $b_1 \leq \frac{\sqrt{2}}{2}(1 + \epsilon - 2\delta)$. For $\delta = 0.01$ and $\epsilon = .001$ we get $b_1 < 0.694$.

Using the penalty method, an estimate $(\hat{\sigma}_s, \hat{\theta}_s)$ is obtained finding the maximum of $\mathrm{f}_{\star,P}(\sigma, \theta) = \mathrm{f}_{\star,P}^{(1)}(\sigma, \theta) + \mathrm{f}_{\star,P}^{(2)}(\sigma, \theta)$.

4.4 Steps 5 and 6

The estimate of the discovery probability at the next iteration, $\hat{U}_{s+0}(0)$, is computed as described in Sect. 3, Eq 4. If its value is lower than $p_\star$ the algorithm stops, otherwise the next iteration $s{+}1$ is performed (if $s{+}1 > M_\star$ the algorithm stops).

5 Conclusion

Given an optimality crierion ϕ, the problem of ϕ-optimal design generation has been addressed. A methodology to support the decision whether to continue or stop the search for optimal designs has been developed. It combines recent advances on discovery probability estimation, based on a Bayesian non parametric approach, Favaro et al (2012), with well known methods for optimal design generation.

In principle, this methodology could be applied to any discrete optimisation problem. This topic will be part of future research.

A software code, written in SAS, that makes use of the Proc Optex procedure, has been developed.

Acknowledgements I would like to thank both Mauro Gasparini (Politecnico di Torino), Gasparini (2012), and Giovanni Pistone (Collegio Carlo Alberto, Moncalieri, Torino) for the helpful discussions I had with them.

References

(2010) SAS/QC(R) 9.2 User's Guide, Second Edition. Cary, NC, United States

Atkinson AC, Donev AN, Tobias RD (2007) Optimum experimental designs, with SAS. Oxford University Press New York

Basso D, Salmaso L, Evangelaras H, Koukouvinos C (2004) Nonparametric testing for main effects on inequivalent designs. In: Bucchianico A, Luter H, Wynn H (eds) mODa 7 Advances in Model-Oriented Design and Analysis, Contributions to Statistics, Physica-Verlag HD, pp 33–40, DOI 10.1007/978-3-7908-2693-7\textunderscore4, URL `http://dx.doi.org/10.1007/978-3-7908-2693-7_4`

Clark JB, Dean A (2001) Equivalence of fractional factorial designs. Statistica Sinica 11(2):537–548

Favaro S, Lijoi A, Prunster I (2012) A new estimator of the discovery probability. Biometrics 68(4):1188–1196, DOI 10.1111/j.1541-0420.2012.01793.x, URL `http://dx.doi.org/10.1111/j.1541-0420.2012.01793.x`

Fedorov VV (1972) Theory of optimal experiments

Gasparini M (2012) Mixtures and limits of symmetric random integer partitions. Metron LXX(2-3):1–11

Giancristofaro RA, Fontana R, Ragazzi S (2012) Construction and nonparametric testing of orthogonal arrays through algebraic strata and inequivalent permutation matrices. Commun Stat, Theory Methods 41(16-17):3162–3178, DOI 10.1080/03610926.2011.579380

Mitchell TJ, Miller Jr F (1970) Use of design repair to construct designs for special linear models. Math Div Ann Progr Rept(ORNL-4661) pp 130–131

Pukelsheim F (2006) Optimal design of experiments, vol 50. Society for Industrial Mathematics

Shah KR, Sinha BK (1989) Theory of optimal designs, vol 582. Springer-Verlag New York

Wynn HP (1970) The sequential generation of d-optimum experimental designs. The Annals of Mathematical Statistics 41(5):1655–1664